STATISTICS IN MEDICINE, VOL. 13, 623-634 (1994)

THE PARTIAL QUESTIONNAIRE DESIGN FOR CASE-CONTROL STUDIES

SHOLOM WACHOLDER

National Cancer Institute, Biostatistics Branch, Epidemiologic Methods Section, 6130 Executive Blvd., EPN 403, Rockville, MD 20852, U.S.A.

RAYMOND J. CARROLL

National Cancer Institute, Biostatistics Branch, Epidemiologic Methods Section, 6130 Executive Blvd., EPN 403, Rockville, MD 20852, U.S.A. and Department of Statistics, Texas A&M University, College Station, TX 77842-3143, U.S.A.

DAVID PEE

Information Management Systems, 6110 Executive Blvd., Rockville, MD 20852, U.S.A.

AND

MITCHELL H. GAIL

National Cancer Institute, Biostatistics Branch, Epidemiologic Methods Section, 6130 Executive Blvd., Rockville, MD 20852, U.S.A.

SUMMARY

We propose an alternative to a long questionnaire that may increase quality while reducing the cost and effort of participants and researchers. In the 'partial questionnaire design', information about the exposure of interest is obtained from all subjects, while zero, one, or more disjoint subsets of questions about possible confounders are asked to randomly selected subgroups. The proposed analyses exploit the fact that the uncollected data can be considered to be missing at random. We show that it is possible to obtain high efficiency for estimating the effect of exposure of interest, adjusted for confounding, while substantially shortening average questionnaire length.

1. INTRODUCTION

A lengthy questionnaire for an epidemiologic study can result in lower rates of participation by potential study subjects, lower quality in those who do participate but become less conscientious with time, added cost for the study, and added burden to participants. In this paper we propose a method we call the 'partial questionnaire design' (PQD) that can reduce the average time needed for completion of a questionnaire with only a minor loss in statistical efficiency compared to the standard method using the same number of participants.

Many epidemiologic studies obtain information in considerable detail about several risk factors that are not themselves the focus of the investigation, but rather are possible confounders or effect-modifiers of the relationship between the exposure of interest and the study disease. In the PQD, each secondary variable is determined for only a fraction of study subjects and subsets of individuals are asked about distinct but overlapping subsets of the study variables. We concentrate on the simplest form of the PQD below. All the secondary variables are split into two

vector-valued variables, denoted as Z_1 and Z_2. Each individual is randomly assigned into one of four categories. All subjects are asked about the exposure of interest X; subjects in category C_{11} are asked about Z_1 and Z_2, in category C_{10} or C_{01} about Z_1 or Z_2, respectively, and in category C_{00} about neither.

We develop methods for analysing a case-control study that uses the partial questionnaire design. Since the investigator randomly determines who will be missing which variables, the data can be missing completely at random (MCAR), or, if the known value of disease status, a demographic factor, or the exposure of interest is allowed to affect the type of questionnaire given to the subject, missing at random (MAR), in the sense of Little and Rubin.[1] We develop MAR methods to allow us to use different missingness probabilities in cases and controls.[1]

The special case of the PQD, in which X and Z_1 are obtained from everyone and Z_2 is obtained on a random subset of participants, can be analysed by methods developed for the two-stage design.[2-11] The variables X and Z_1 are collected in the first stage and a subset of participants are studied further in the second stage to obtain Z_2. It is straightforward[1] to extend this example to the more general case of *monotone missingness*, where the I covariates can be ordered in a way that whenever the ith covariate is missing, so too are covariates $i+1, i+2, \ldots, I$. However, methods of analysis different from those previously proposed for two-stage designs are needed for the general PQD problem because some subjects will be missing both Z_1 and Z_2, some will be missing only one covariate, and some both, resulting in *non-monotone missingness*.[1] We estimate the parameters in a prospective risk model by applying an estimating equation method. Our approach can handle non-monotone missingness and needs only a small proportion of subjects with complete data in order to get high efficiency.

We first outline methods of analysis (Section 2) and then study the relative efficiencies of various PQD allocations for a realistic example based on a case-control study of risk factors for oral cancer[12] (Section 3). Simulations confirm that the estimation procedure we propose yields well-behaved point estimates and confidence intervals (Section 3). The limitations of our results and possible extensions are discussed in Section 4.

2. METHODS

2.1. Estimation of parameters

We call the exposure of interest X and the secondary variables in the two subsets that can be missing Z_1 and Z_2, respectively. We define the four categories of missingness and the measured covariates and their likelihood contributions in Table I. These categories apply to both cases ($d = 1$) and controls ($d = 0$).

Suppose disease incidence in the base or source population satisfies the prospective logistic risk model

$$Pr(D = 1|X, Z_1, Z_2; \beta^*) \equiv \frac{\exp(Z'\beta^*)}{1 + \exp(Z'\beta^*)} \equiv H(\beta^*; X, Z_1, Z_2) \tag{1}$$

where $Z' = (1, X, Z_1, Z_2)$ and $\beta^{*'} = (\beta_0^*, \beta_1, \beta_2, \beta_3)$. Applying Bayes' theorem to the population of cases and controls in the case-control sample (see Mantel[13] and Prentice and Pyke[14]), one finds that there is a new intercept β_0 such that

$$f(x, z_1, z_2|d) = \frac{H(\beta; x, z_1, z_2)^d \{1 - H(\beta; x, z_1, z_2)\}^{1-d} q(x, z_1, z_2)}{Pr(D = d)} \tag{2}$$

Table I. Sampling categories for cases and controls

Category	Indicator	Measured Covariates	Likelihood contribution
C_{11}	$\Delta(C_{11})$	X, Z_1, Z_2	$f(x, z_1, z_2 \mid d)$
C_{01}	$\Delta(C_{01})$	X, Z_2	$f(x, z_2, \mid d)$
C_{10}	$\Delta(C_{10})$	X, Z_1	$f(x, z_1 \mid d)$
C_{00}	$\Delta(C_{00})$	X	$f(x \mid d)$

in the case-control population. In equation (2), $\beta' = (\beta_0, \beta_1, \beta_2, \beta_3)$; $q(X, Z_1, Z_2)$ is the mass function of X, Z_1 and Z_2 *in the case-control sample*; and $Pr(D = 1)$ is the proportion of cases *in the case-control sample*. Thus, assuming Z_2 is missing at random,

$$f(x, z_1 \mid D) = \frac{\sum_{z_2} H(\beta; x, z_1, z_2)^d \{1 - H(\beta; x, z_1, z_2)\}^{1-d} q(x, z_1, z_2)}{Pr(D)}. \tag{3}$$

Other conditional probabilities needed to construct the likelihood are obtained similarly. A case in category C_{10} with $X = x$ and $Z_1 = z_1$ contributes $f(x, z_1 \mid D = 1)$. Other subjects contribute similar factors, depending on their case-control status and missingness category. Letting $\Delta_i(C_{jk}) = 1$ if the ith individual is in category C_{jk} and zero otherwise, the factor contributed to the likelihood $\mathscr{L}$ by an individual with case-control status $D = d_i$ is

$$\begin{aligned}
&\{H(\beta; X, Z_1, Z_2)^{d_i} [1 - H(\beta; X, Z_1, Z_2)]^{1-d_i} q(X, Z_1, Z_2)\}^{\Delta_i(C_{11})} \\
&\quad \times \left\{ \sum_{z_1} H(\beta; X, z_1, Z_2)^{d_i} [1 - H(\beta; X, z_1, Z_2)]^{1-d_i} q(X, z_1, Z_2) \right\}^{\Delta_i(C_{01})} \\
&\quad \times \left\{ \sum_{z_2} H(\beta; X, Z_1, z_2)^{d_i} [1 - H(\beta; X, Z_1, z_2)]^{1-d_i} q(X, Z_1, z_2) \right\}^{\Delta_i(C_{10})} \\
&\quad \times \left\{ \sum_{z_1, z_2} H(\beta; X, z_1, z_2)^{d_i} [1 - H(\beta; X, z_1, z_2)]^{1-d_i} q(X, z_1, z_2) \right\}^{\Delta_i(C_{00})}.
\end{aligned} \tag{4}$$

We estimate β and q by finding the unconstrained maximum of $\mathscr{L}$. As noted by Prentice and Pyke,[14] a maximum likelihood procedure would maximize this likelihood subject to the constraint that $Pr(D = 1)$ equals the proportion of cases in the case-control sample. Even though our procedure does not necessarily satisfy this constraint in small samples, the unconstrained score equations, obtained by differentiation of log $\mathscr{L}$ with respect to β and q, have expectation zero and thus lead to parameter estimates $\hat{\beta}$ and $\hat{q}$ that are consistent and asymptotically normal. These results, which are described elsewhere by Carroll *et al.*,[15] apply to discrete-valued exposures and covariates or to continuous covariates if one is willing to postulate a parametric model for q. This theory has not yet been developed for continuous covariates and non-parametric estimates of q.

Interactions between X and Z_1 can be analysed using the likelihood (4) by defining

$$H(\cdot) \equiv \frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 X Z_1)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 X Z_1)}. \tag{5}$$

For discrete exposures and covariates, convenient starting values for maximizing $\mathscr{L}$ are obtained by the Expectation Maximization (EM) algorithm.[16] The E-step calculates expected numbers of

cases and controls for cells defined by levels of X, Z_1 and Z_2, conditional on all the data and on the estimates of β and q from the previous M-step. The M-step fits the regression model with standard software for complete data, as if the frequencies generated in the E-step were complete data from the study.

Once good starting values are obtained, Newton–Raphson iteration based on analytic first and second derivatives can be used to accelerate convergence. A quasi-Newton procedure based on numerical differentiation to obtain first and second derivatives[17] as implemented in GAUSS 2.1 (Aptech Systems; Kent, Washington, 1991) yielded virtually the same results.

2.2. Estimation of the covariance and relative efficiency

An analysis of the score equations based on log $\mathscr{L}$ reveals that, while the overall expectation is zero, each case and control does not contribute a mean zero component to the score. Consequently, the variances of the parameter estimates $\hat{\beta}$ are, in theory, smaller than obtained from the inverse of the matrix of second derivatives of log $\mathscr{L}$.[15] Nevertheless, numerical studies indicate that the correction term is often negligible. Hence, in most applications suitably accurate covariance estimates can be obtained from the inverse of the hessian of log $\mathscr{L}$.[15]

The relative efficiency for estimates of a given parameter under various designs is obtained as the ratio of theoretical variances. For these calculations, the more precise variance formulae of Carroll *et al.*[15] were used.

3. EFFICIENCY OF VARIOUS DESIGNS HYPOTHETICALLY APPLIED TO A CASE-CONTROL STUDY OF ORAL CANCER

3.1. Description of the data

We investigated the properties of the partial questionnaire design based on data from a case-control study of the effect of smoking on oral and pharyngeal cancer.[12] We thank Dr. William J. Blot for his permission to use this data set as an example. Key scientific goals of the study included estimating the effect of smoking and evaluating possible modification of the smoking effect by drinking of alcohol. Alcohol consumption and number of missing teeth were regarded as possible confounders.

We used subsets of the study subjects to simulate the PQD in a study with a sample size more typical of studies of cancer etiology. First, we sampled 600 controls and 200 cases randomly with replacement from the 1108 cases and 1264 controls with known values of the variables of interest in the original study. The distributions of cases and controls and the log-odds estimates and their standard errors from the full questionnaire design (FQD) applied to these 800 subjects are presented in Tables II and III. Using the notation in equation (5), we note that the confounding effect of Z_1 ($\beta_2 = 1{\cdot}34$) is much stronger than that of Z_2 ($\beta_3 = -0{\cdot}081$): the corresponding odds ratio relating Z_1 and disease (3·82) represents a much stronger relationship than that of Z_2 and disease (odds ratio of 0·92). Both Z_1 and Z_2 have strong associations with X (odds ratio of 2·34 for Z_1 and 2·50 for Z_2). We fit two models, one with only the three main effects parameters ($\beta_1, \beta_2, \beta_3$) of dichotomous X (20 or more years duration of cigarette smoking), Z_1 (15 or more drinks per week), and Z_2 (7 or more lost teeth), and the other also including a parameter (β_4) for the X by Z_1 interaction. We compared the relative efficiencies for $\beta_1, \beta_2, \beta_3$ and β_4 in various types of PQD against the FQD in which all 800 subjects provided the full data. The properties of the parameter and variance estimators and the estimated relative efficiencies appear to be adequate based on a simulation study described in the next section.

Table II. Joint distribution of X, Z_1, and Z_2 in controls for test data set using random subset of original data of Blot *et al.*[12]

	Controls				Cases			
	$X=1$		$X=0$		$X=1$		$X=0$	
	$Z_1=1$	$Z_1=0$	$Z_1=1$	$Z_1=0$	$Z_1=1$	$Z_1=0$	$Z_1=1$	$Z_1=0$
$Z_2=1$	54	85	37	138	66	30	9	15
$Z_2=0$	21	48	39	178	36	17	15	12

Table III. Parameter and variance estimates of log-odds ratios based on full questionnaire design (FQD) using random subset of data of Blot *et al.*[12]

Parameter	Parameter and variance estimate	
	No-interaction model	$X \times Z_1$ interaction model
β_1	1·46	1·44
	0·038	0·070
β_2	1·34	1·31
	0·034	0·095
β_3	− 0·081	− 0·081
	0·036	0·036
β_4	—	0·043
		0·15

3.2. Effect of changing design parameters

The parameter and variance estimates from a single random realization of the PQD and the expected efficiency for various design matrices for the PQD are shown in Table IV. Panel 1 of Table IV is the FQD analysed as described in connection with equation (5) for a PQD. As expected, the parameter and variance estimates based on maximizing equation (5) are the same as for standard logistic regression. We believe that the relative efficiencies below 100 per cent reflect the small price of estimating q.

In panel 2, cases and controls are assigned equally to each of the four categories in Table I, resulting in a 50 per cent reduction in the numbers of subjects with measured Z_1 values and a 50 per cent reduction in those with measured Z_2 values. In the main effects model, the loss of efficiency in estimating the effect of X is only 11 per cent; for Z_1 and Z_2 the loss is slightly over 50 per cent. In the interaction model, there is a loss of half of the efficiency in estimating the interaction.

The design in panel 3, with only 20 cases and controls in each categories C_{11}, C_{01} and C_{10} and all other subjects in category C_{00} could produce a great reduction in average questionnaire length since no information on covariates is obtained from 85 per cent of subjects and only one of Z_1 or Z_2 is obtained from 10 per cent. The loss of efficiency in the main effects model for estimating the effect of X is less than 50 per cent. However, the losses for estimating other effects are quite substantial; thus the designs in panel 3 would be appropriate only if there were no interest in the effects of secondary covariates or in interactions.

Table IV. Parameter estimate, variance estimate, and 100 times the relative efficiency for several PQDs in a subset of the oral-pharyngeal cancer data[12]

Panel	Disease	Number in design category					Main effects model		$X \times Z_1$ interaction model				
		C_{11}	C_{01}	C_{10}	C_{00}		β_1	β_2	β_3	β_1	β_2	β_3	β_4
1	$d=0$	600	0	0	0	$\hat{\beta}$	1·46	1·34	−0·08	1·44	1·31	−0·08	0·04
	$d=1$	200	0	0	0	s.e.*	0·20	0·18	0·19	0·27	0·31	0·19	0·38
	Total	800	0	0	0	R.E.†	100	100	98	100	100	98	100
2	$d=0$	150	150	150	150	$\hat{\beta}$	1·52	1·46	−0·29	1·65	1·63	−0·26	−0·27
	$d=1$	50	50	50	50	s.e.*	0·21	0·26	0·28	0·35	0·44	0·28	0·55
	Total	200	200	200	200	R.E.†	89	50	45	66	50	45	50
3	$d=0$	20	20	20	540	$\hat{\beta}$	1·12	1·95	−0·03	0·79	1·34	−0·25	1·03
	$d=1$	20	20	20	140	s.e.*	0·31	0·56	0·58	0·50	0·89	0·50	1·21
	Total	40	40	40	680	R.E.†	52	12	12	28	15	12	13
4	$d=0$	120	120	120	120	$\hat{\beta}$	1·64	1·69	−0·52	2·09	2·27	−0·42	−0·88
	$d=1$	40	40	40	80	s.e.*	0·24	0·30	0·34	0·43	0·52	0·34	0·64
	Total	160	160	160	320	R.E.†	84	40	36	56	40	36	40
5	$d=0$	300	0	0	300	$\hat{\beta}$	1·36	1·26	0·32	1·48	1·44	0·32	−0·25
	$d=1$	100	0	0	100	s.e.*	0·20	0·26	0·27	0·32	0·44	0·27	0·54
	Total	400	0	0	400	R.E.†	89	50	49	66	50	49	60
6	$d=0$	150	180	120	150	$\hat{\beta}$	1·49	1·46	0·02	1·03	0·87	0·07	0·90
	$d=1$	50	60	40	50	s.e.*	0·21	0·27	0·27	0·35	0·46	0·27	0·58
	Total	200	240	160	200	R.E.†	88	45	49	62	45	49	45
7	$d=0$	150	120	180	150	$\hat{\beta}$	1·46	1·56	−0·18	1·29	1·34	−0·20	0·35
	$d=1$	50	40	60	50	s.e.*	0·21	0·25	0·30	0·32	0·42	0·30	0·52
	Total	200	160	240	200	R.E.†	89	55	41	70	55	41	55
8	$d=0$	114	162	162	162	$\hat{\beta}$	1·42	1·21	0·10	1·16	0·82	0·12	0·60
	$d=1$	50	50	50	50	s.e.*	0·21	0·26	0·28	0·32	0·45	0·29	0·56
	Total	164	212	212	212	R.E.†	88	48	43	65	49	43	49
9	$d=0$	123	159	159	159	$\hat{\beta}$	1·52	1·30	−0·10	1·43	1·17	−0·09	0·20
	$d=1$	41	53	53	53	s.e.*	0·21	0·27	0·29	0·34	0·45	0·29	0·56
	Total	164	212	212	212	R.E.†	87	47	42	63	47	42	47
10	$d=0$	150	150	150	150	$\hat{\beta}$	1·46	1·74	−0·17	1·77	2·14	−0·12	−0·63
	$d=1$	14	62	62	62	s.e*	0·23	0·29	0·34	0·40	0·51	0·34	0·62
	Total	164	212	212	212	R.E.†	85	41	36	55	40	36	41

* Standard error
† Relative efficiency

The effect of assigning 40 per cent of subjects to category C_{00} and 20 per cent to the other three categories is shown in panel 4. The relative efficiency of 84 per cent for the main effect of X is, as expected, greater than for the design in panel 3. Figure 1 displays how the relative efficiency for estimating β_1 varies with the proportions assigned to C_{00} when the other subjects are assigned equally to the other three categories.

The design in panel 5 divides subjects equally between C_{11} and C_{00} and has the same number of subjects asked about Z_1 and Z_2 as does panel 2. The relative efficiency for the main effect of X is slightly higher in panel 5 (89 per cent versus 88·6 per cent). Designs intermediate between those in panels 2 and 5 have relative efficiencies that increase monotonically from 88·6 to 89·1 per cent. These result support the conjecture of Zelen[18] that the most efficient design for fixed

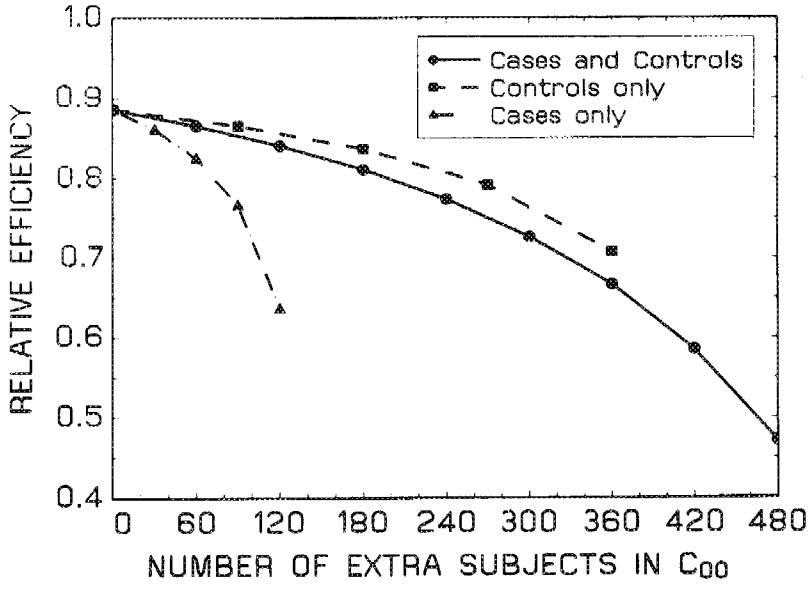

Figure 1. Impact of moving equal numbers of subjects from C_{10}, C_{01} and C_{11} to C_{00} on the relative efficiency for estimating the main effect of X under risk model 1, for a subset of the data from Blot *et al.*[12] The abscissas represent differences between the *numbers* of subjects in C_{00} and the number of subjects in C_{00} in the 'benchmark' design where 25 per cent of cases and controls are in each category (panel 2 of Table IV). The ordinates are the relative efficiencies compared to the FQD. The locus of squares describes the effect of moving only controls from C_{10}, C_{01} and C_{11} to C_{00}; the locus of triangles describes the effect of moving only cases; and the locus of circles describes the effect of moving one case for each three controls

numbers of subjects asked about Z_1 and Z_2 assigns subjects only to categories C_{11} or C_{00}. However, in this example all these designs have similar efficiency.

In this data set, the efficiency is not affected substantially by switching subjects from category C_{01} to category C_{10}, even though the confounding effect of Z_1 is much greater, as described in connection with Tables II and III. In panels 6 and 7, one can see the effect of changing the numbers of subjects missing Z_1 or Z_2. The relative efficiency for β_1 in the main effect model is higher in panel 2, with equal numbers missing Z_1 and Z_2, than in panels 6 and 7. These differences are small, however, suggesting that the balanced design is reasonable, even when one covariate is a stronger confounder than the other. Figure 2 displays this phenomenon over a broader range of designs.

In these studies we had three times as many controls as cases. Relative efficiency for the main effect and the interaction is more sensitive to changing the category distribution of cases than of controls, and intermediate for the mixture of both; this is as expected since the estimates of β_2 and β_3 will be more precise when Z_1 and Z_2 are observed by proportionately more cases and there are more controls than cases in the study.[19] In panels 8, 9 and 10 of Table IV, the total number of subjects is the same. The relative efficiencies for estimation of all seven parameters decreased monotonically as the number of cases with complete information decreased. Table IV as well as all the figures suggest that obtaining more complete information from a higher proportion of cases than controls gives more efficiency for fixed total numbers of subjects in each category.

These results suggest that the PQD provides an opportunity for substantial savings if Z_1 or Z_2 is expensive to obtain. Use of fractions of 0·25 for each of the four categories results in high

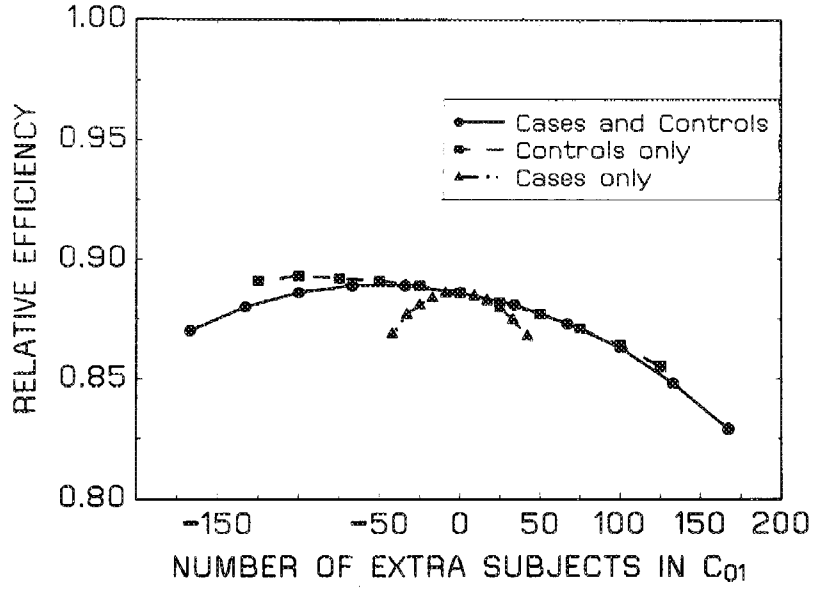

Figure 2. Impact of moving subjects from C_{10} to C_{01} on the relative efficiency for estimating the main effect of X under risk model 1, for a subset of the data from Blot *et al.*[12] The abscissas represent differences between the *numbers* of subjects in C_{01} and the number of subjects in C_{01} in the 'benchmark' design where 25 per cent of cases and controls are in each category (panel 2 of Table IV). The ordinates are the relative efficiencies compared to the FQD. The locus of squares describes the effect of moving only controls from C_{10} to C_{01} the locus of triangles describes the effect of moving only cases, and the locus of circles describes the effect of moving one case for each three controls

efficiency and a 50 per cent reduction in the marginal effort to obtain Z_1 and Z_2, but even lower proportions in the categories C_{11}, C_{01} and C_{10} do not result in major deterioration of efficiency.

The efficiency loss for assessing interactions between X and Z_1 or Z_2 is, as expected, similar to the efficiency loss for estimating the main effect of Z_1 or Z_2. The reduction in precision of estimates of these interactions, in contrast to the loss in efficiency in estimating β_1, can be substantial (Table IV and Figure 3) and is usually close to the percentage of subjects for whom the covariate was not measured. Therefore, we do not recommend obtaining only partial information on variables whose interactions with exposure are of interest, particularly since the precision of estimates of interactions often is low even for the FQD in typical case-control studies.[20]

4. SIMULATION STUDY

We simulated hypothetical studies to examine whether point and interval estimates based on the estimation procedures in Section 2 had nominal operating characteristics for typical sample sizes. We examined whether the theoretical variance discussed in Section 2 differed noticeably from an estimate based simply on the hessian of log $\mathscr{L}$. Simulations were also used to determine whether efficiency calculations based on the expected information provide useful guidance for deciding whether or not to use the PQD and, if so, exactly what design parameters to employ.

In a simulation study of 1000 replications of the PQD with 500 controls and 200 cases, we let $\beta_1 = \beta_2 = 1$ and $\beta_3 = 1{\cdot}5$. For controls, the joint distribution of the binomial variables X, Z_1 and Z_2 was the multinomial distribution obtained from a log-linear model with all parameters (other than the mean) equal to zero, except for the parameters of the XZ_1 and XZ_2 interactions that

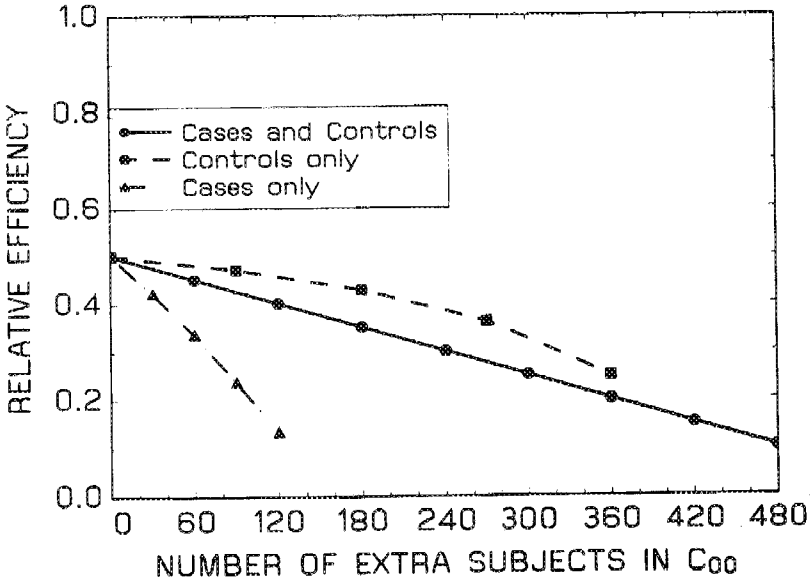

Figure 3. Impact of moving equal numbers of subjects from C_{10}, C_{01} and C_{11} *to* C_{00} on the relative efficiency for estimating the interaction between X and Z_1 under risk model 5, for a subset of the data from Blot *et al.*[12] The abscissas represent differences between the *numbers* of subjects in C_{00} and the number of subjects in C_{00} in the 'benchmark' design where 25 per cent of cases and controls are in each category (panel 2 of Table IV). The ordinates are the relative efficiencies compared to the FQD. The locus of squares describes the effect of moving only controls from C_{10}, C_{01} and C_{11} to C_{00}; the locus of triangles describes the effect of moving only cases; and the locus of circles describes the effect of moving one case for each three controls

were set equal to log (2). The joint distribution of X, Z_1 and Z_2 in cases can be calculated from the control distribution and β_1, β_2, β_3. In the studies, 20, 25, 25, and 30 per cent of controls and cases were in categories C_{11}, C_{01}, C_{10} and C_{00}, respectively. Applying the correction of Carroll *et al.* reduced the variance estimates by about one part in 5000. For the PQD, the 95 per cent confidence interval based on the uncorrected variance covered the true values of β_1 961 times, which is within the 95 per cent limits of 936–964 for the number of successes from 1000 independent Bernoulli experiments with proportion of success of 0·95. The ratio of the empirical variance to the variance calculated from the expected information was 1·063, which falls within the interval 0·91–1·09 that contains the ratio of empirical variance from a sample of size 1000 to the true variance for a normal variate with a probability of 0·95. Other simulation studies, including some using interaction models, other numbers of subjects, other values of β and other joint distributions of the exposure and covariates, suggest that the PQD estimators we present have good operating characteristics (data not shown).

5. DISCUSSION

Our work has two important limitations. We only considered the simple situation of dichotomous X, Z_1 and Z_2. It is not clear that the small sample and efficiency properties of the PQD are as good as those we found when the dimension of $X \times Z_1 \times Z_2$ increases or when more covariates are involved. It is also unclear how to extend the methods of analysis to continuous covariates without invoking parametric models. Second, the results in Table IV are based on

a single data example. It would be useful to study other examples, including some with more extreme confounding effects of Z_1 and Z_2.

Our results do suggest that the PQD can be implemented with small loss of efficiency for estimation the main effect of the exposure X. Investigators have flexibility in choosing the parameters of the design in ways that can reduce the cost to investigators and the burden to participants, as long as the proportion of subjects for whom complete information is obtained is not too low. Efficiency loss is substantial, however, for estimation of the main effect of Z_1 or Z_2 or of interactions like XZ_1.

These efficiency results are not surprising. One can consider an estimate of β_1 in the no-interaction model as the sum of a crude effect based on the relation of X alone with disease and the logarithm of the confounding risk ratios[19] of Z_1 and Z_2. It has been shown theoretically[19,21,22] and demonstrated empirically[23,24] that, except in extreme situations, the adjusted odds ratio for exposure is not very sensitive to the strength of the confounder-disease or confounder-exposure association. Therefore, the precision of the adjusted estimate should not be greatly affected even when there is substantial variability in the estimates of the parameters contributing to the confounding risk ratio. Thus, for example, in comparing panels 2, 6 and 7 of Table IV, or in examining Figure 2, the impact of switching subjects from missing Z_1 to missing Z_2 appears to be small. For estimating the interaction, on the other hand, the loss of efficiency is, as expected, approximately proportional to the numbers of subjects who do not have all the variables involved in the interaction.

It is possible, of course, not to collect any information on Z_1 (or, equivalently, on Z_2), reducing questionnaire length even further. Then β_1 would be the estimate of the effect of X-adjusted for Z_2 alone from the FQD. However, this design does not yield a consistent estimator of β_1 unless $\beta_2 = 0$. When it is clear that the confounding risk ratios for Z_1 and Z_2 are close to unity, the savings from not collecting one or more secondary covariate may overshadow any possible bias. The PQD might be considered when the possibility of important bias is considered to be less remote or when adjustment for Z_1 or Z_2 is required for the credibility of the study.

Simulations indicate that asymptotic theory leads to valid point and interval estimates and to good estimates of the relative efficiency of various possible designs in samples of moderate size.

Field studies are needed to determine whether a higher participation rate can be achieved by asking potential subjects to submit to the shorter PQD questionnaire. An increase in participation could overcome some of the reduction in statistical efficiency of a PQD. Use of computer-assisted interviewing[25] could handle the logistics of matching the subject to the appropriate questionnaire, with minimal burden on the interviewer.

A crucial requirement for our methods of analysis is that the missing data be missing at random. This assumption may be violated if the chance of participation depends on the length of the interview that the subject is asked to complete. One approach to avoid this problem is to tell all subjects about the design in advance and consider only those who agree to accept *any* assigned questionnaire as eligible for the studies. The missing at random assumption could also be violated if there is a reduction in the quality of responses as the interview proceeds. Violations of this assumption could have implications for the FQD as well as the PQD.

One special PQD deserves attention. Our studies suggest that the design in which some of the subjects answer all questions and the others only provide data on X can have high statistical efficiency for estimating the main effect. This design can be regarded as a two-stage design in which the first stage consists of measuring X on all cases and controls and the second stage in measuring other covariates on a subset of subjects. A further advantage of this design is that the analytical methods of Breslow and Cain[5] can be used. These methods are applicable to continuous covariates; however, the exposure of interest must be discrete in this application. Although

the two-stage design has several attractive features, it does not reduce the numbers of subjects who must answer the full questionnaire by as much as some other PQD designs that are almost as efficient.

In the PQD subjects are either asked or not asked about a covariate. An alternative approach would be to either ask about the covariates in full detail or in a brief question. For example, one could randomly assign subjects to be asked either for a detailed smoking history or for a yes–no answer to a simple question, such as, 'Have you ever been a regular smoker?' This alternative could recapture some of the efficiency lost in the PQD with very little extra effort.

The techniques in this paper have other applications. In a study with several exposures of interest, one exposure, X, may require a larger sample size to achieve estimates with the desired precision than other exposures, say Z_1 and Z_2. Since adjustment for all other exposures may be desirable, the PQD could be used by measuring X on everyone, and Z_1 and Z_2 on overlapping subsamples. For example, the PQD is being considered for a prospective study of the effects of pesticide exposure, of diet and cooking practices, and of physical activity of cancer risk. In this study, the desired sample size for the pesticide component (X) may be greater than the sample size required for study of diet and cooking practices (Z_1) and physical activity (Z_2).

We believe that the PQD offers potential practical advantages, such as increased participation, that can outweigh the small loss of statistical efficiency. However, additional work is required to handle continuous exposures and covariates and to study efficiency over a broader range of parameter values. Field studies are needed both to determine whether the PQD yields studies with higher participation and data quality and to determine whether or not the assumption of 'missing at random' is tenable in practice.

ACKNOWLEDGEMENT

Research by one author (R.J.C.) supported by a grant from the National Cancer Institute.

REFERENCES

1. Little, R. J. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
2. Walker, A. M. 'Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known', *Biometrics*, **38**, 1025–1032 (1982).
3. White, J. E. 'A two stage design for the study of the relationship between a rare exposure and a rare disease', *American Journal of Epidemiology*, **115**, 119–128 (1982).
4. Fears, T. R. and Brown, C. C. 'Logistic regression methods for retrospective case-control studies using complex sampling procedures', *Biometrics*, **42**, 955–960 (1986).
5. Breslow, N. E. and Cain, K. C. 'Logistic regression for two-stage case-control data', *Biometrika*, **75**, 11–20 (1988).
6. Breslow, N. E. and Zhao, L. P. 'Logistic regression for stratified case-control studies', *Biometrics*, **44**, 891–899 (1988).
7. Weinberg, C. R. and Wacholder, S. 'The design and analysis of case-control studies with biased sampling', *Biometrics*, **46**, 963–975 (1990).
8. Zhao, L. P. and Lipsitz, S. 'Designs and analysis of two-stage designs', *Statistics in Medicine*, **11**, 769–782 (1990).
9. Flanders, W. D. and Greenland, S. 'Analytic methods for two stage case-control studies and other stratified designs', *Statistics in Medicine*, **10**, 739–747 (1991).
10. Scott, A. J. and Wild, A. J. 'Fitting logistic regression models in stratified case-control studies', *Biometrics*, **47**, 479–510 (1991).
11. Wacholder, S. and Weinberg, C. R. 'Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling', *Biometrics*, in press.
12. Blot, W. J., McLaughlin, J. K., Winn, D. M., Austin D. M., Greenberg R. S. Preston-Martin, S., Bernstein, L., Schoenberg, J. B., Stemhagen, A. and Fraumeni, J. F. 'Smoking and drinking in relation to oral and pharyngeal cancer', *Cancer Research*, **48**, 3282–3287 (1988).

13. Mantel, N. 'Synthetic retrospective studies and relative topics', *Biometrics*, **29**, 479–486 (1973).
14. Prentice, R. L. and Pyke R. 'Logistic disease incidence models and case-control studies', *Biometrika*, **66**, 403–411 (1979).
15. Carroll, R. J., Wang S. and Wang C. Y. 'Asymptotics for prospective analysis of stratified case-control studies', *Journal of the American Statistical Association*, in press (1994).
16. Dempster, A. P., Laird N. M. and Rubin, D. B. 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B*, **39**, 1–38 (1977)
17. Dennis, J. E., Jr. and Schnabel, R. B. *Numerical methods for unconstrined optimization and nonlinear equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
18. Zelen, M. 'Discussion', *Statistics in Medicine*, **13**, 647-649 (1994).
19. Breslow, N. E. and Day, N. E. *Statistical methods in cancer research, Vol 1, The analysis of case-control studies (IARC Scientific Publications No. 32)*, International Agency for Research on Cancer, Lyon, 1980.
20. Greenland, S. 'Tests for interaction in epidemiologic studies: a review and study of power', *Statistics in Medicine*, **2**, 243–251 (1983).
21. Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. 'Smoking and lung cancer: Recent evidence and a discussion of some questions', *Journal of the National Cancer Institute*, **22**, 173–203 (1959).
22. Yanagawa, T. 'Case-control studies: assessing the effect of a confounding factor', *Biometrika*, **71**, 191–194 (1984).
23. Blair, A., Hoar, S., Walrath, J. 'Comparison of crude and smoking-adjusted standardised mortality ratios', *Journal of Occupational Medicine*, **27**, 881–884 (1885).
24. Siemiatycki, J., Wacholder, S., Dewar, R. *et al.* 'Degree of confounding bias related to smoking, ethnic group, and socioeconomic group in estimates of the associations between occupation and cancer', *Journal of Occupational Medicine*, **30**, 617–625 (1988).
25. Saris, W. E. *Computer-Assisted Interviewing. Sage University Paper Series on Quantiative Applications in the Social Sciences, 07-080*, Sage, Newbury Park, CA, 1991.